# Learning from Big Data

Hendrik Blockeel

**Industry 4.0 - How to get there**
**Leuven, Oct 19, 2016**

With contributions from Jesse Davis,
and Luc De Raedt, and Wannes Meert

DTAI
KU LEUVEN

# Some background…

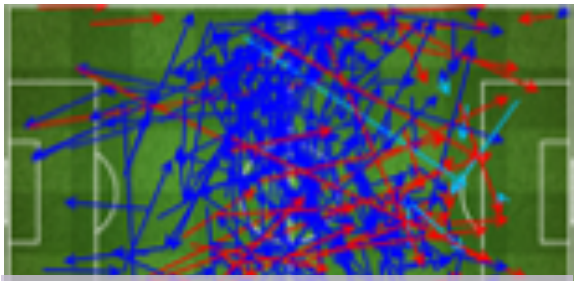
- Research group “Declarative languages and artificial intelligence” (DTAI), Dept. Computer Science, KU Leuven
  - 10 Faculty
  - ±13 Postdoctoral Researchers
  - ±40 PhD Researchers
- Large subgroup working on Machine Learning

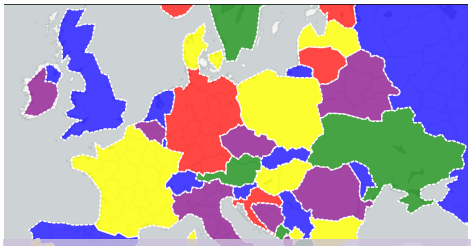
DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

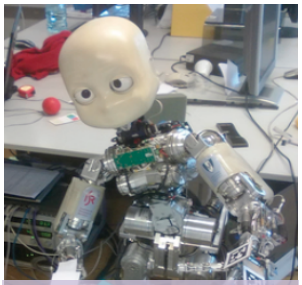
# Research @ DTAI — KU Leuven

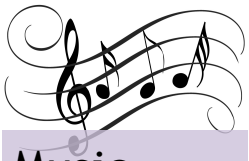

Bio-informatics

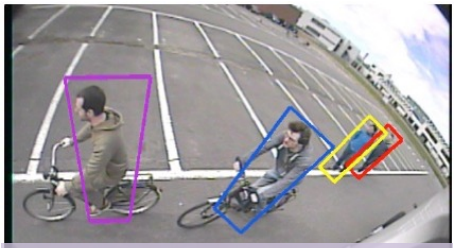

Sports Analytics


Knowledge Base Systems

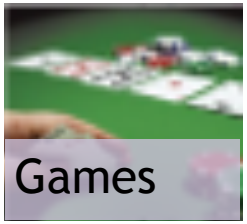

Chem-informatics

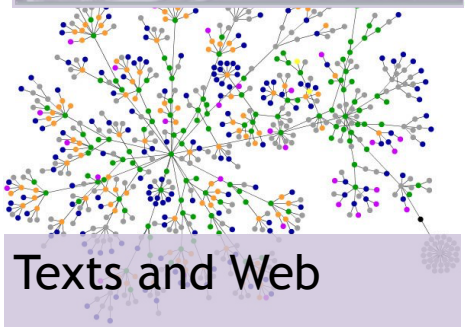

Robotics


Music


Computer Vision


Health


Games


Texts and Web


Engineering and Sensors

https://dtai.cs.kuleuven.be/topics

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# What connects all these things?

Central in much of this research:

## Machine learning

a.k.a.: data mining, data analytics, **data science**, …

[Cleveland 2001]

# Some examples of “data science”

# Predicting ship arrivals

Metro,
April 24, 2014

**Big data kan haven efficiënter maken**

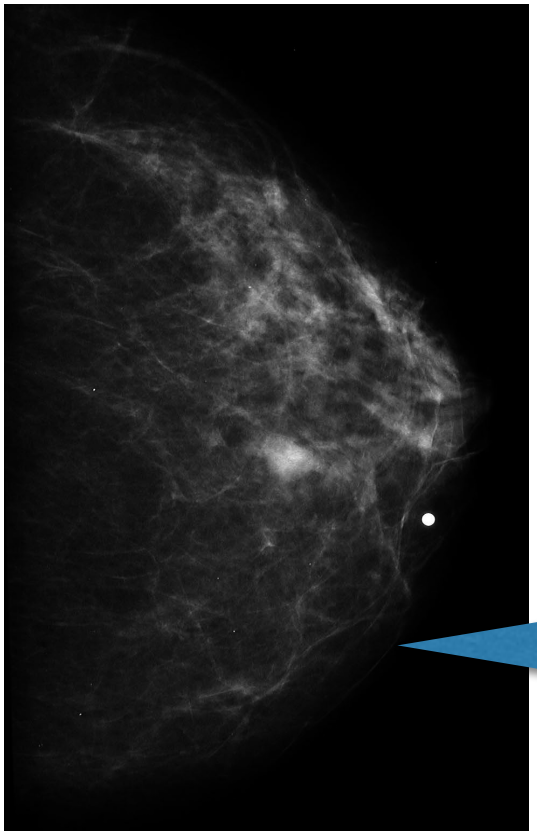
Foto Belga

**BRUSSEL** Door onvoorziene omstandigheden is het soms moeilijk het exacte aankomstmoment van een schip te berekenen, wat ook gevolgen heeft voor de terminalactiviteiten. Uit onderzoek van de Universiteit Antwerpen blijkt dat het toepassen van het zogenaamde 'Random Forest'-algoritme op verschillende soorten informatie vertragingen het nauwkeurigst kan inschatten. Dat is een algoritme waarin gegevens over aankomsttijden, scheepstypes en vorige havens werden gebundeld om zo preciezer het aankomstuur van een schip te bepalen. Die kennis kan de logistieke afhandeling van goederen efficiënter en dus ook goedkoper maken, klinkt het. ■

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Medical diagnosis



Abnormality in mammogram: benign or malignant?
*Learn*, from data, a model that predicts this.

Courtesy of Jesse Davis, KU Leuven

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Square Kilometre Array

- www.skatelescope.org/
- To be ready in 2024
- largest radio telescope ever
- “will produce 10 times the global internet traffic”

- *How to analyze all that?*

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# InfraWatch, “Hollandse brug”

continuous monitoring of bridge
to avoid another urgent close-down

Courtesy of Joost Kok, Leiden University

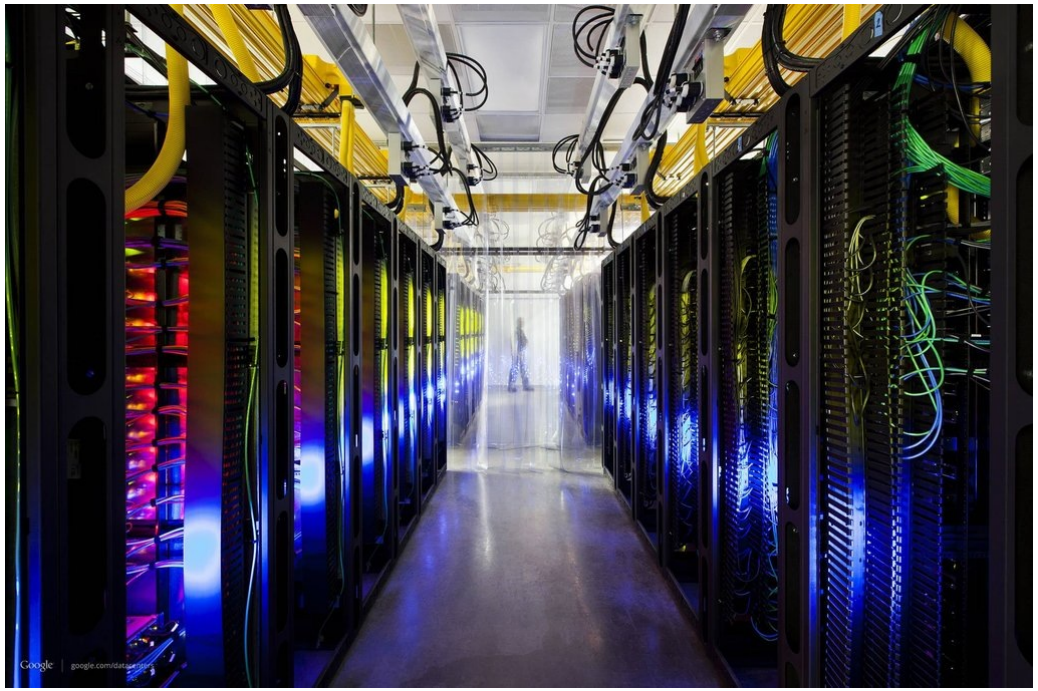
# Predictive maintenance

- Example: In-flight data collection in airplanes
- Lots of data collected, sometimes anomalies (error messages, warning lights, …)
- Search for “root cause” during maintenance: takes time
- Automatic analysis of in-flight data may help

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Data Centers



Google data center

NSA data center, Utah

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# What is Data Science?

"The next new hype?"
Data Analytics … Data Mining … Statistics … Prognostics …
Predictive Analytics … Machine learning … Big Data …

A natural definition: **Data Science is the science that studies all aspects of the process of collecting and exploiting data.**

So, a bit like statistics … (which is 200 years old) ?

[Cleveland 2001]

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# It's not *just* statistics

- Statistics does not tell you …
  - how to store terabytes of data so that you can efficiently retrieve what you're looking for
  - how to organize complex data
- That's Databases…

- Statistics also doesn't tell you …
  - how to index data stored on millions of computers and make it available as if it were on your hard disk
  - how to send terabytes of data from a telescope to a data center
- That's Distributed Systems…

- Statistics won't help you out if
  - rigid mathematical methods cannot be used (necessary assumptions don't hold)
  - difficulties are algorithmic rather than mathematical
- That's Machine learning…

- Statistics does not focus on handling
  - enormous volumes of data
  - arriving at high speed
  - of a complex nature
- That's Data mining …

DTAI
Declaratieve Talen en Artificiële Intelligentie

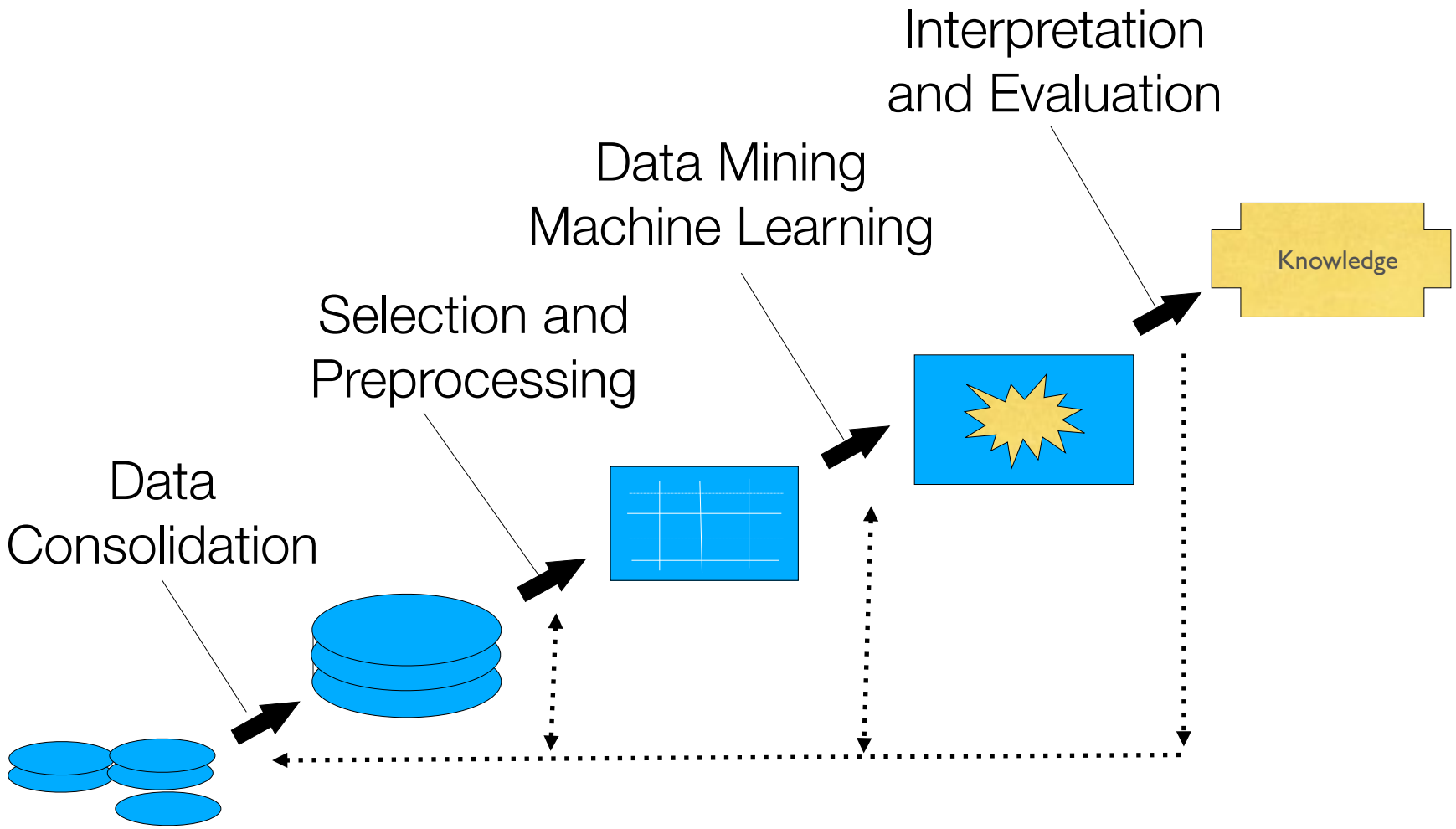
KU LEUVEN

# Data Science according to Wikipedia

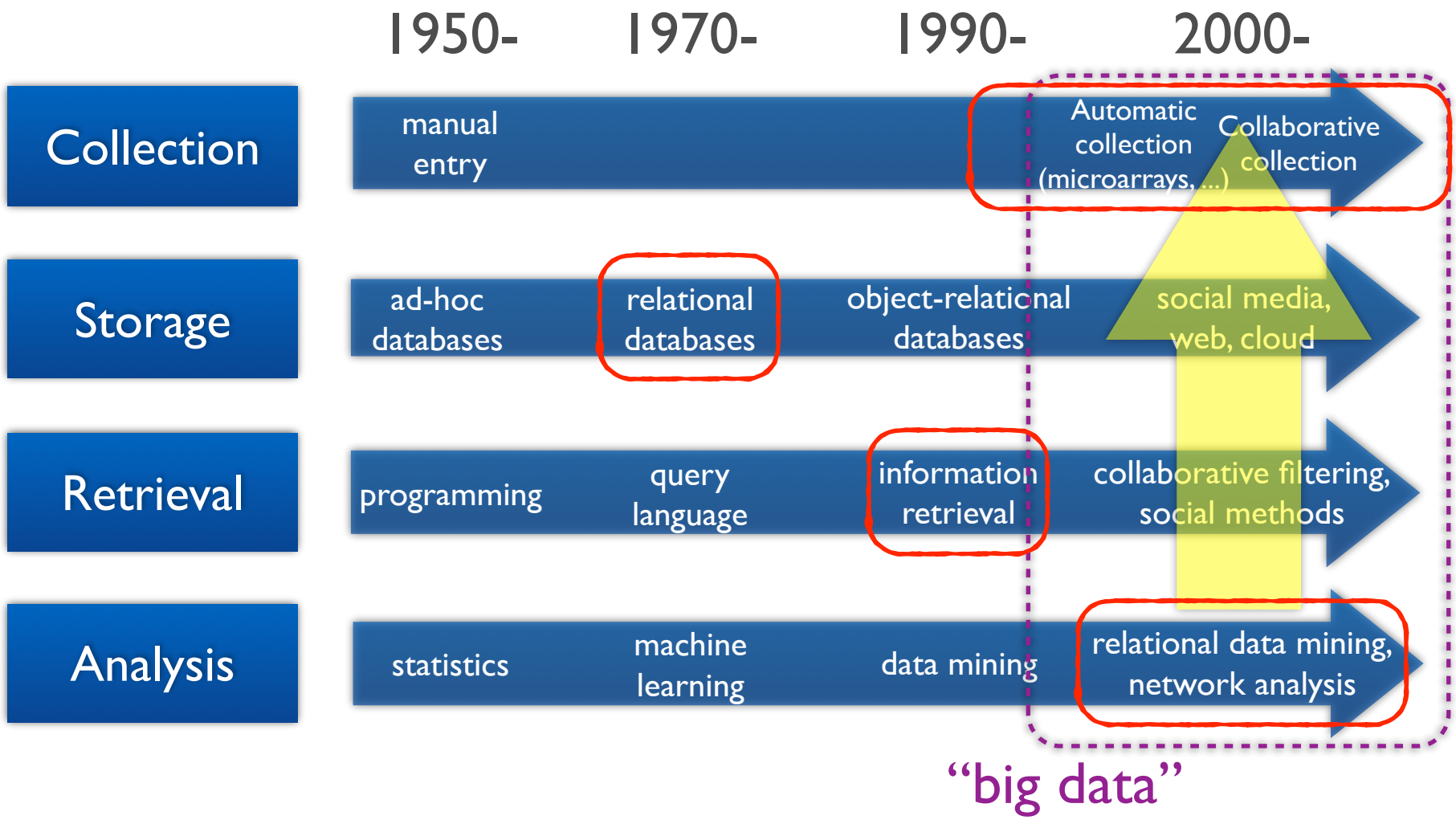
Data science is the study of the generalizable extraction of knowledge from data, yet the key word is science. It incorporates varying elements and builds on techniques and theories from many fields, including signal processing, mathematics, probability models, machine learning, statistical learning, computer programming, data engineering, pattern recognition and learning, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products. The subject is not restricted to only big data, although the fact that data is scaling up makes big data an important aspect of data science. Another key ingredient that boosted the practice and applicability of data science is the development of Machine Learning - a branch of Artificial Intelligence - which is used to uncover patterns from data and develop practical and usable predictive models.

Wikipedia, Sept. 2014

# The Knowledge Discovery Process



# Historical overview




DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# What is Data Mining?

- Data mining is the extraction of patterns and models from data
- Knowledge Discovery and Data mining (KDD) is the overall process of selecting and preparing data, selecting projections, selecting data mining methods, extracting patterns, evaluating patterns as potential knowledge and consolidating knowledge
- Data mining is one step in the KDD process. It is the automated step. Some would call it the machine learning step.

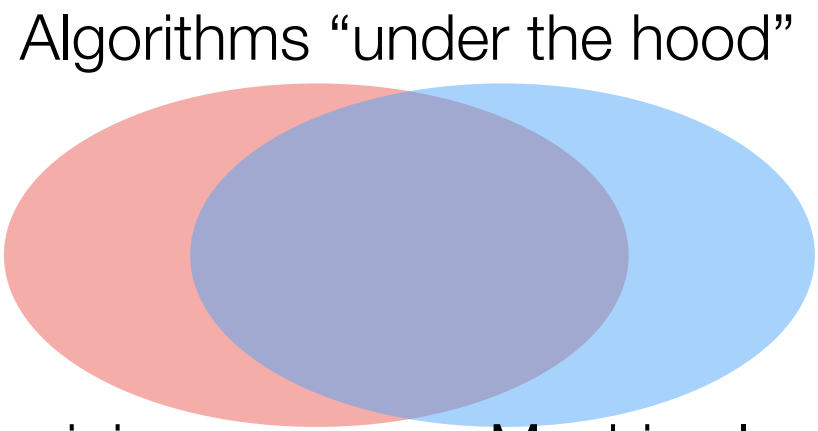
  - *Pattern* is local (describes some data points)
    E.g. all students in the DM course possess data.
  - *Model* is global (describes all data points)
    E.g. predict for all students whether they possess data or not.

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# What is Machine Learning?

- A machine learns if it is able to
  - improve its performance
  - on a specific task
  - with experience
- Closely related to data mining, as experience needs to be analyzed in order to learn




DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# What is Big Data?

- The “four Vs” of Big Data
  - High Volume

    Scale of data
  - High Velocity

    Analysis of streaming data
  - High Variety

    Heterogeneous data
  - High Veracity

    Uncertainty of data
- Big data concerns not just analysis of the data, but everything around it: collection, storage, transfer, visualization and analysis Synonym for Data Science?

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Big Data

- Used interchangeably
- Subtle difference for some:
  - “Collecting Does Not Mean Discovering"
  - Data Science looks to create models that capture the underlying patterns of complex systems, and codify those models into working applications.
  - Big Data looks to collect and manage large amounts of varied data to serve large-scale web applications and vast sensor networks.

[KDnuggets 2015]

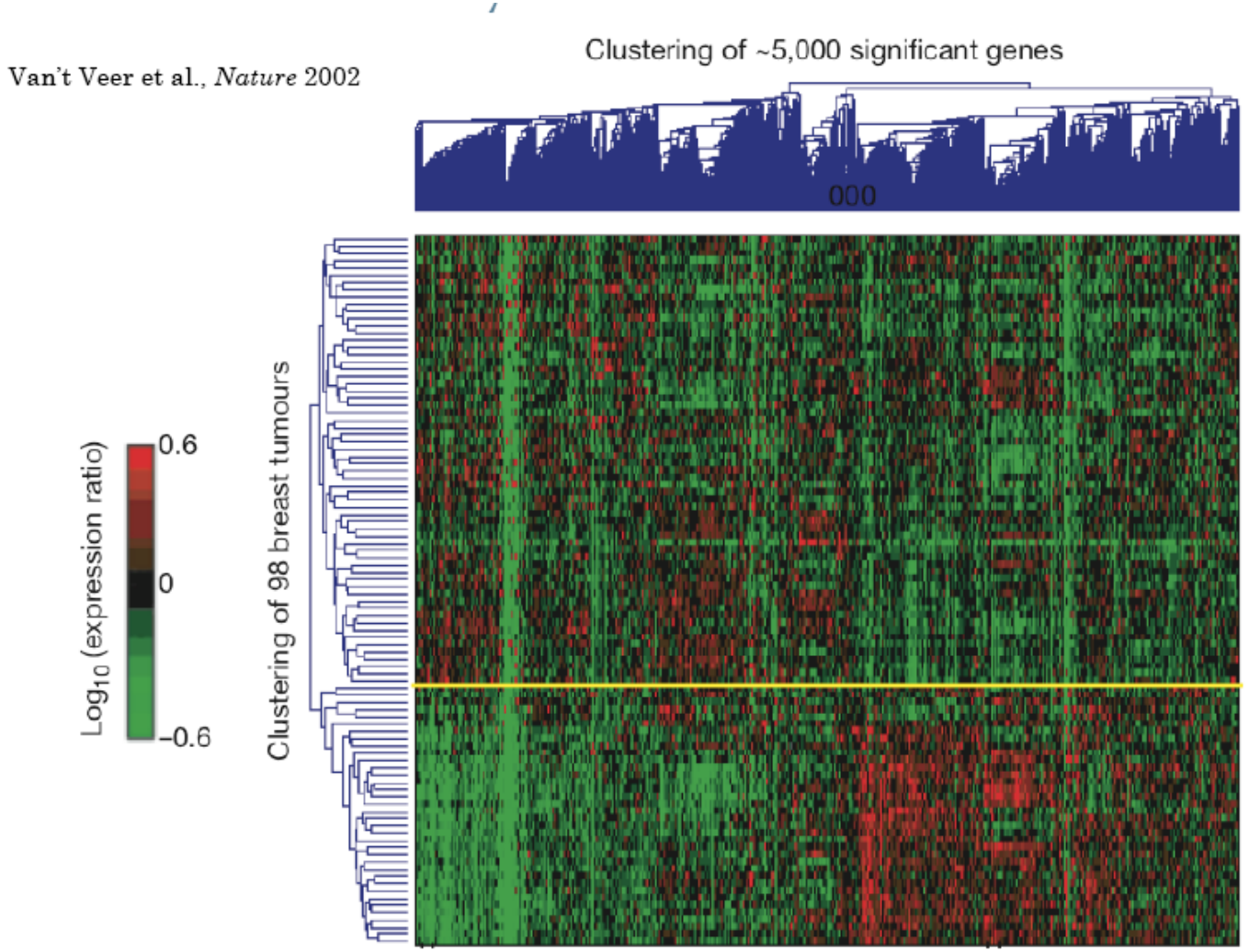
# Machine Learning: tasks and methods

KU LEUVEN

# Learning a model

- In general, machine learning methods learn a **model** of the data
- Once we have that model, we can use it for many things
  - predicting the value of an unknown variable
  - classifying cases into different types
  - detecting instances that do not seem consistent with the model: outliers, or anomalies
  - identifying structure in the data

# Cluster Analysis




DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Associations

Which products are frequently bought together ?

4/5=80%

4/5=80%

4/5=60%

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Predictive learning

- Given: a data set, with elements of the form (x,y)
  - x is the “input”
  - y is the “output”
- Task: **Learn to predict y from x**
- In the operational phase, the system will get examples with x only, and need to predict the corresponding y

- = learn a function f: x → y

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Variants of predictive learning

- Supervised: the target value is known
  - Predict class: classification
  - Predict numbers: regression
  - Predict rank: ranking (or interest)
- Unsupervised: the target value is unknown
  - Group similar objects: clustering
  - Identify “abnormal” cases: anomaly detection
- Semi-supervised: some target values are known

Different settings require different approaches

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Interpretable vs. black-box

The **predictive function** or **model** learned from the data may be represented in a format that we can easily interpret, or not

Non-interpretable models are also called **black-box** models

In some cases, it is crucial that predictions can be explained (e.g.: banks deciding whether to give you a loan)

Note difference between explaining a *model* and explaining a *prediction*

# Supervised, semi-supervised, unsupervised learning

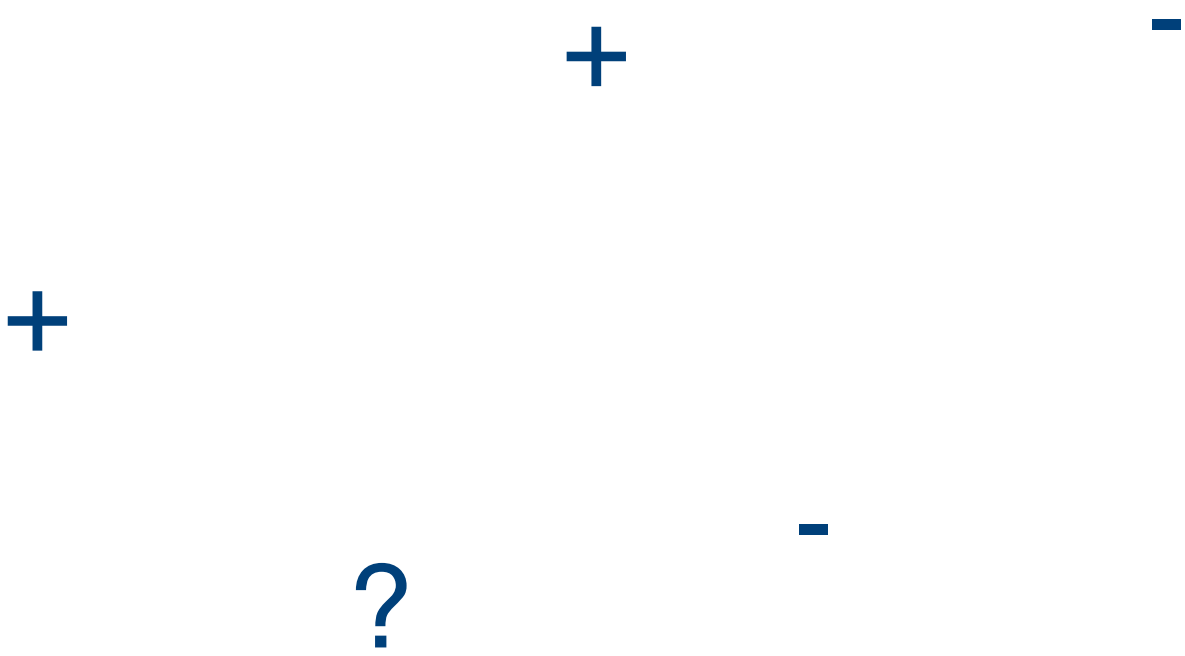
- **Supervised learning**: learning a (predictive) model from *labeled* instances

- **Unsupervised learning**: learning a model from *unlabeled* instances
  - such models are often not predictive (without *any* information on what to predict, how could you learn from that?)
  - yet, may be (indirectly) useful for predictive learning, or for other types of learning

- **Semi-supervised learning**: learn a predictive model from *a few labeled* and *many unlabeled* examples

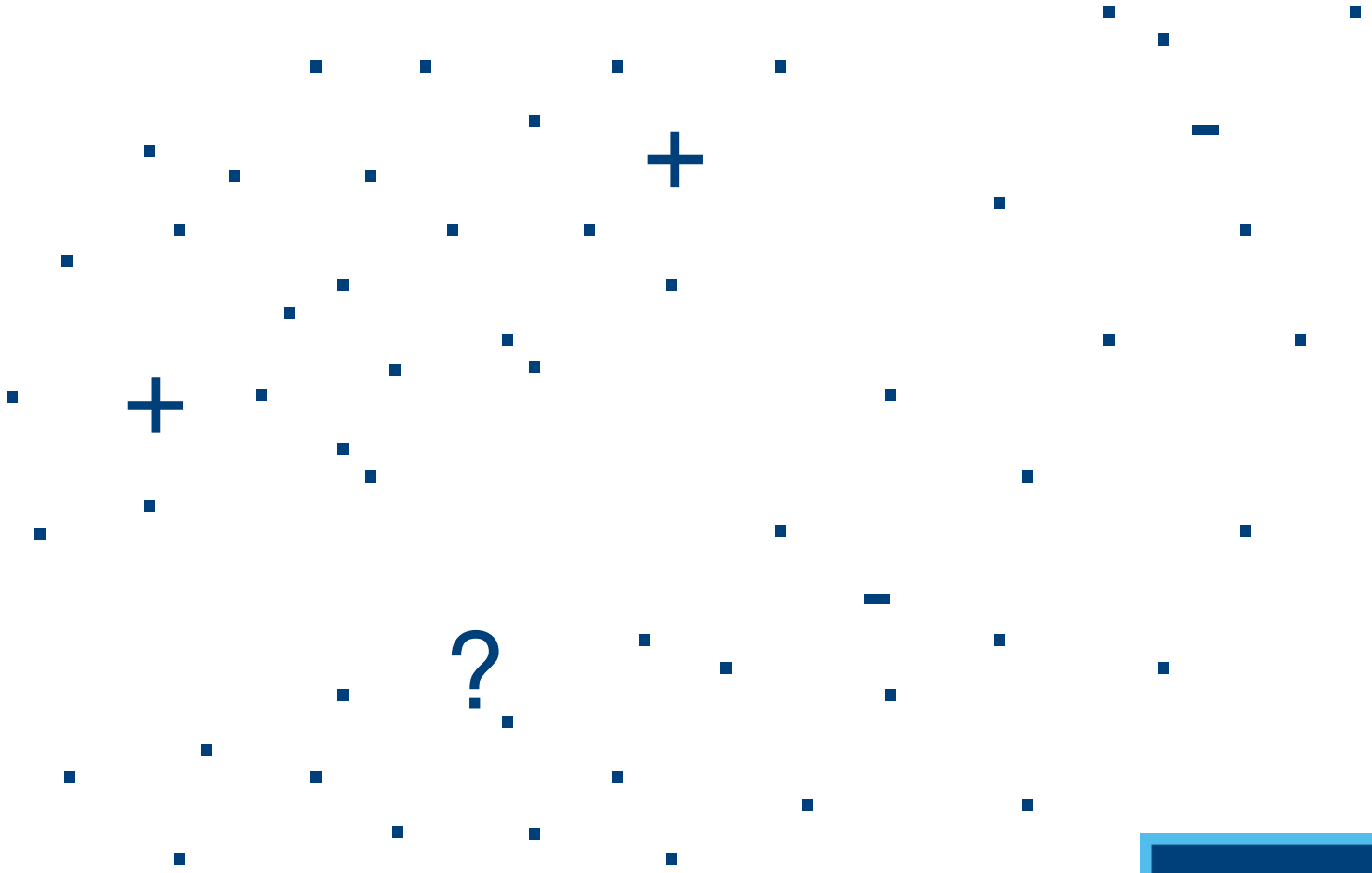
# Semi-supervised learning

- How can unlabeled examples help learn a better model?



# Semi-supervised learning

- How can unlabeled examples help learn a better model?

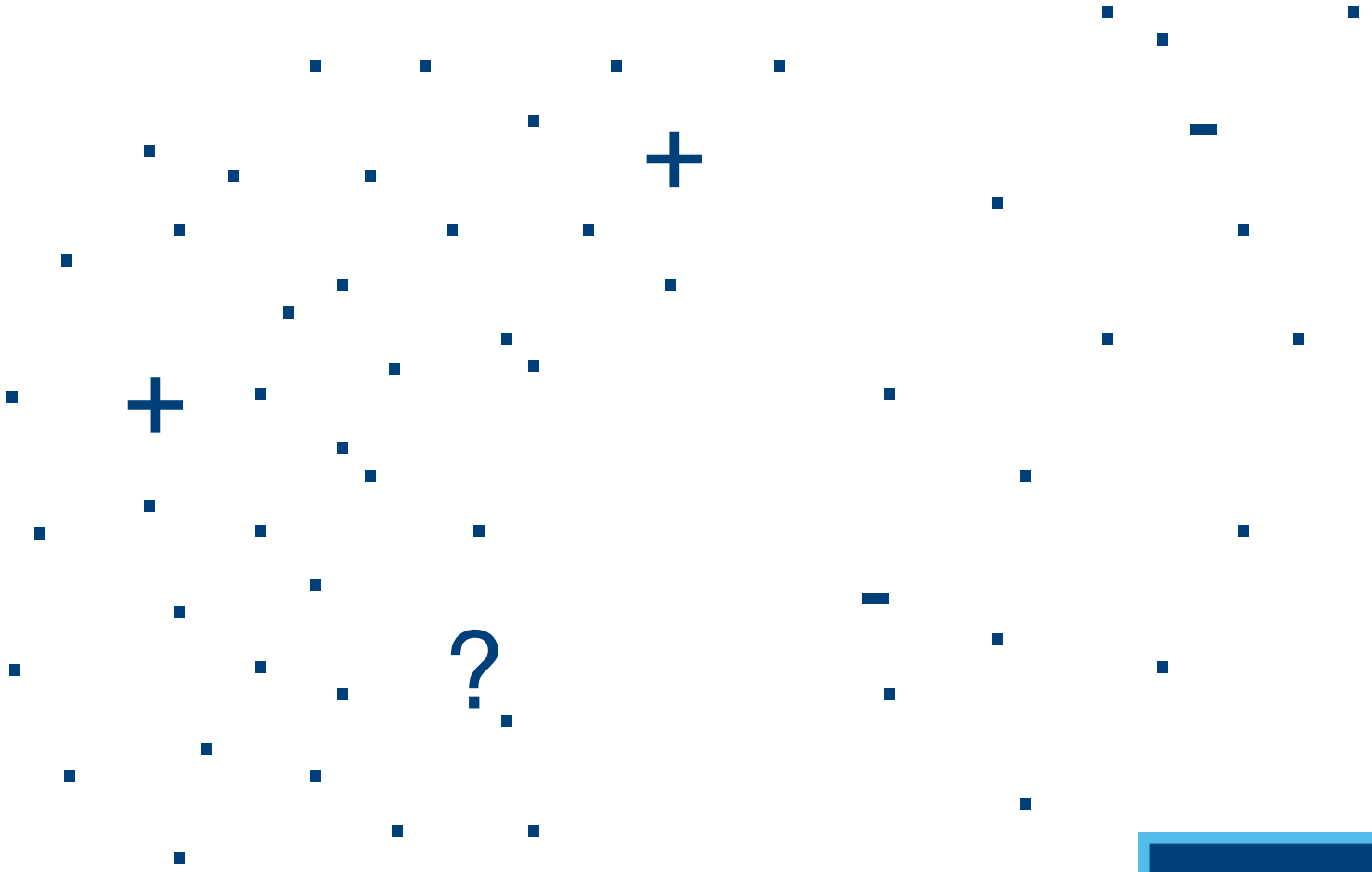



DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Semi-supervised learning

- How can unlabeled examples help learn a better model?




DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Format of input data

- Input is often assumed to be a *set* of instances that are all described using the same variables (features, attributes)
  - The data are “i.i.d.”: “independent and identically distributed”
  - The training set can be seen as a random sample from one distribution
  - The training set can be shown as a table (instances x variables) : *tabular data*
  - This is also called the *standard setting*

- There are other formats: instances can be
  - nodes in a graph
  - graphs (not a vector of number)
  - elements of a sequence
  - …

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Format of input data

Training set

| Sepal length | Sepal width | Petal length | Petal width | Class |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | Setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | Setosa |
| 7.0 | 3.2 | 4.7 | 1.4 | Versicolor |
| 6.3 | 3.3 | 6.0 | 2.5 | Virginica |

Prediction set

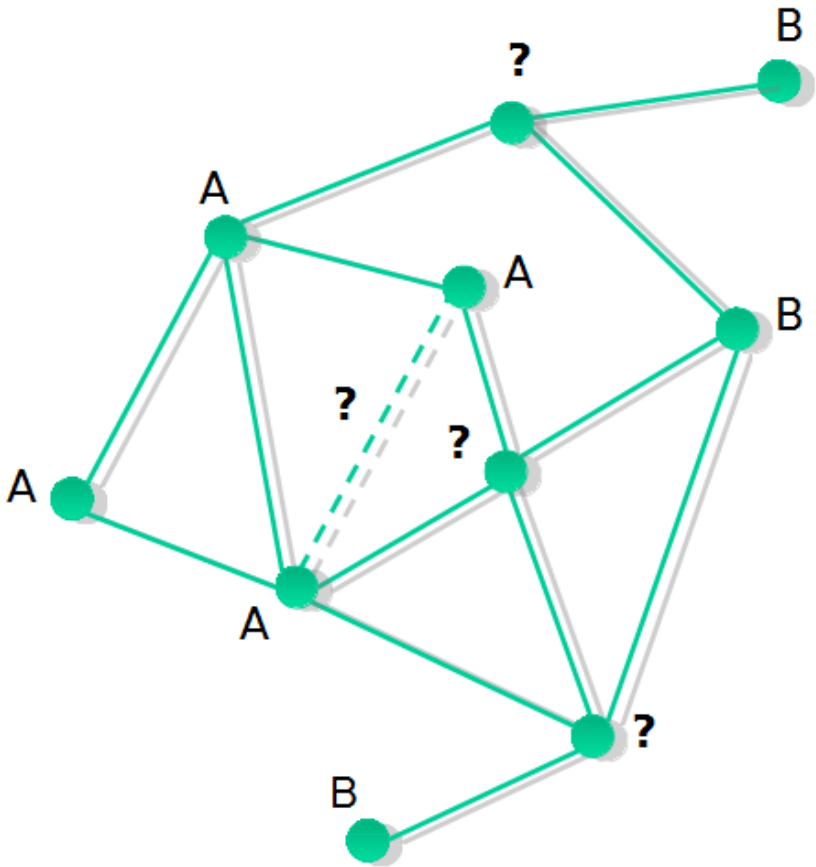
| Sepal length | Sepal width | Petal length | Petal width | Class |
|---|---|---|---|---|
| 4.8 | 3.2 | 1.3 | 0.3 | ? |
| 7.1 | 3.3 | 5.2 | 1.7 | ? |

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Graph data

- Example: Social network
- Target value known for some nodes
- Predict node label
- Predict edge
- Predict edge label
- …
- Use *network structure* for these pre

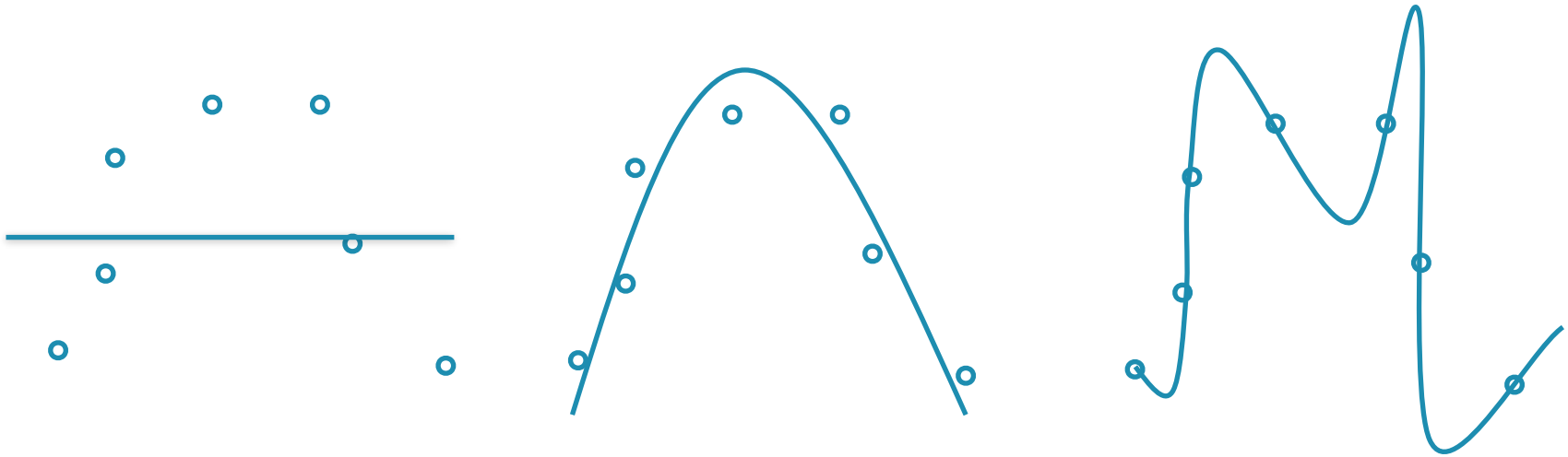



DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Overfitting and underfitting

- Occam's razor: If choice, the simpler model is better
- Trade-off: explain data vs. simplicity
- Both overfitting and underfitting are harmful
- *Models must be tested on a data set that does not overlap with the dataset used for training the model*

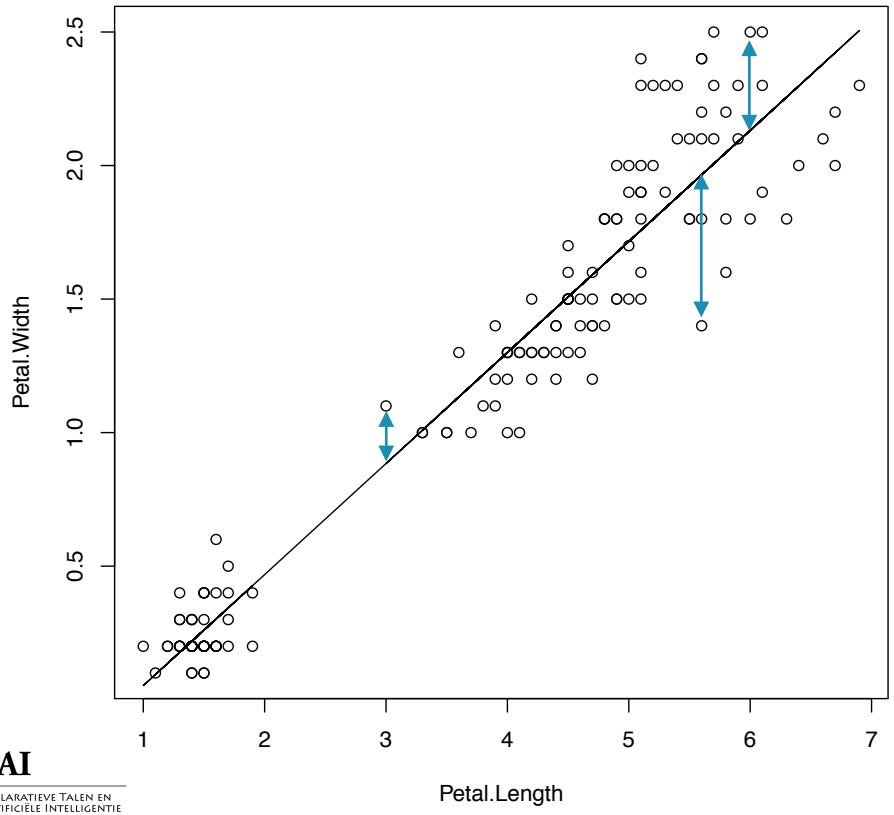


DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Method: Linear regression

- "Linear model" : $Y = a + b_1X_1 + b_2 X_2 + \ldots + b_k X_k$
- Usually fit such that sum of squared vertical deviations from line is minimal ("least squares" method)




to predict petal *width* from petal *length* for iris flowers:

best possible linear predictor (in the sense of "least squares") is

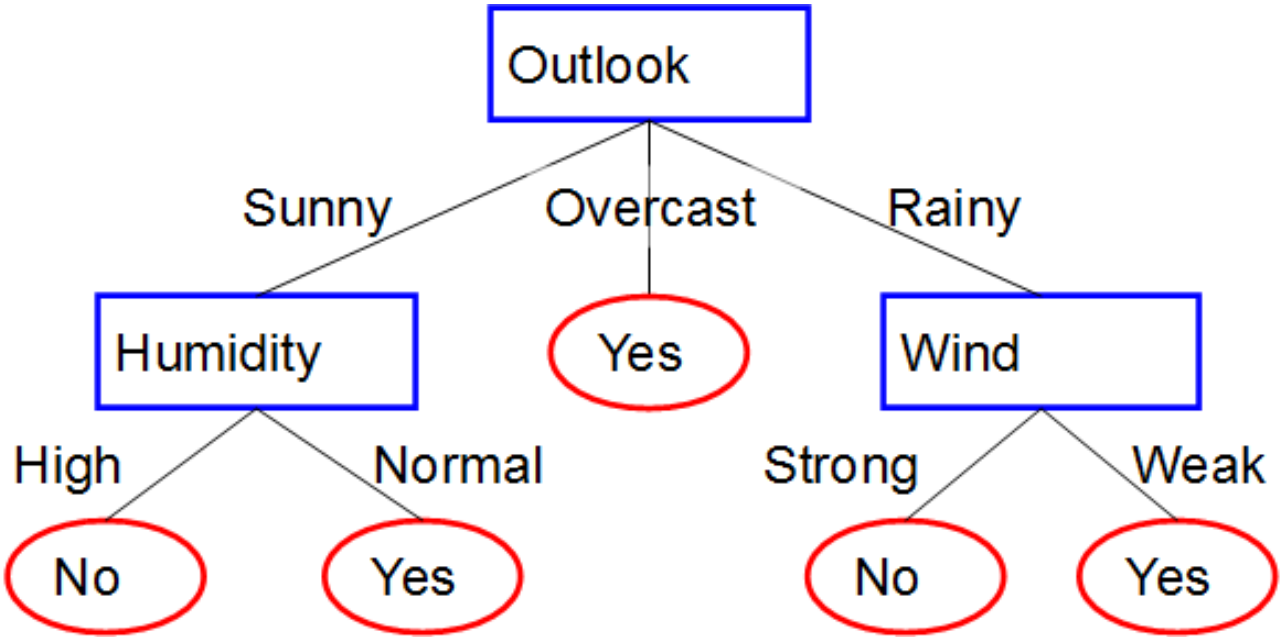
*width = 0.416*length - 0.363*

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Method: decision tree learning

Play tennis or not (depending on weather)

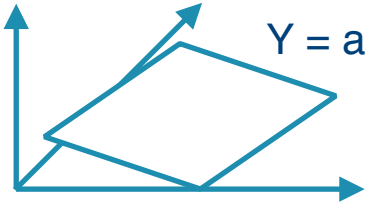



- Internal nodes ask questions
- Leaf nodes contain the prediction
- Very different type of model!

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

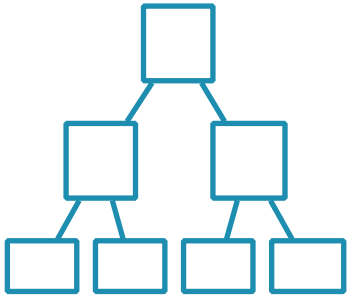
# Trees vs. linear regression



$$Y = a + b_1A + b_2S + b_2C + b_3J$$

On average, professor salaries are 500$ higher than postdoc salaries. I'll assume the same holds for each country separately.

Nah… I don't think so. Until very strong evidence to the contrary is presented, I'll stick to this assumption.



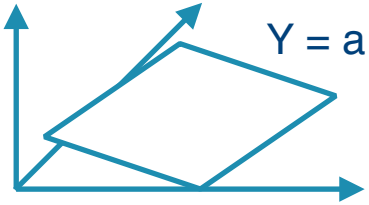
That's silly! You don't know that. It may be different in different countries.

Wow… you're quite biased, you know.

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

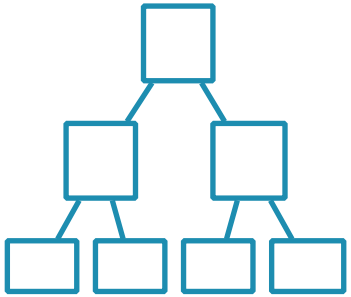
# Trees vs. linear regression



$$Y = a + b_1A + b_2S + b_2C + b_3J$$

Well, surely that'll be the same, no?

I think you're a bit paranoid…



Professors earn 500$ more than postdocs if we look at males aged 30-40; males aged 40-50; males aged 50-60; females aged 30-40; females aged 50-60. But what about females aged 40-50? I have no data!

I'll believe that when I see it with my own eyes, not before.

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Trees vs. linear regression

- Which method will perform best largely depends on your problem
- Each learning approach has a “bias”: implicit assumptions it makes
- Learners whose bias fits your problem will perform better
- Unfortunately, problems are often not sufficiently understood to decide in a principled way (neither are learners, in fact!)

- This holds not just for trees vs. linear models, but for learning methods in general
- ⇒ try different methods !

# Method: Random Forests (RF)

- Learn many decision trees from variants of the same dataset
  - Motivation: “many experts know more than one”
  - One tree may model accidental patterns in the data
  - In the ensemble of trees, non-accidental patterns will dominate

- *Random Forests is one of the most popular predictive methods*
  - easy to use - do not require deep knowledge of the technique
  - often highest predictive accuracy among all competing methods

- Reason for “always high accuracy” : successfully combines bias of “linear” and “tree” models

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Method: Learning rule sets

- Model in the form of “if-then” rules
- Very easy to interpret
- Ideal when relatively simple symbolic models are desired

- E.g., from a set of examples of leap years and non-leap years, could learn the following model:

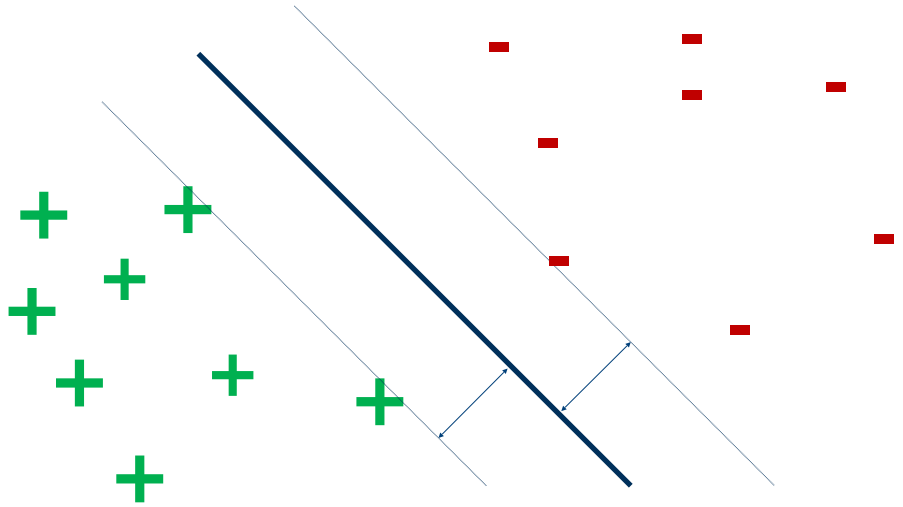
```
IF multiple of 400 THEN leap
ELSE IF multiple of 100 THEN no leap
ELSE IF multiple of 4 THEN leap
ELSE no leap
```

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Method: Support vector machines (SVMs)

- A principled mathematical approach based on convex optimization
- Based on 4 principles: construct a **maximal-margin** linear separator in a high-dimensional **feature space** that is implicitly defined using a **kernel function**, and represent this separator by means of its "**support vectors**"

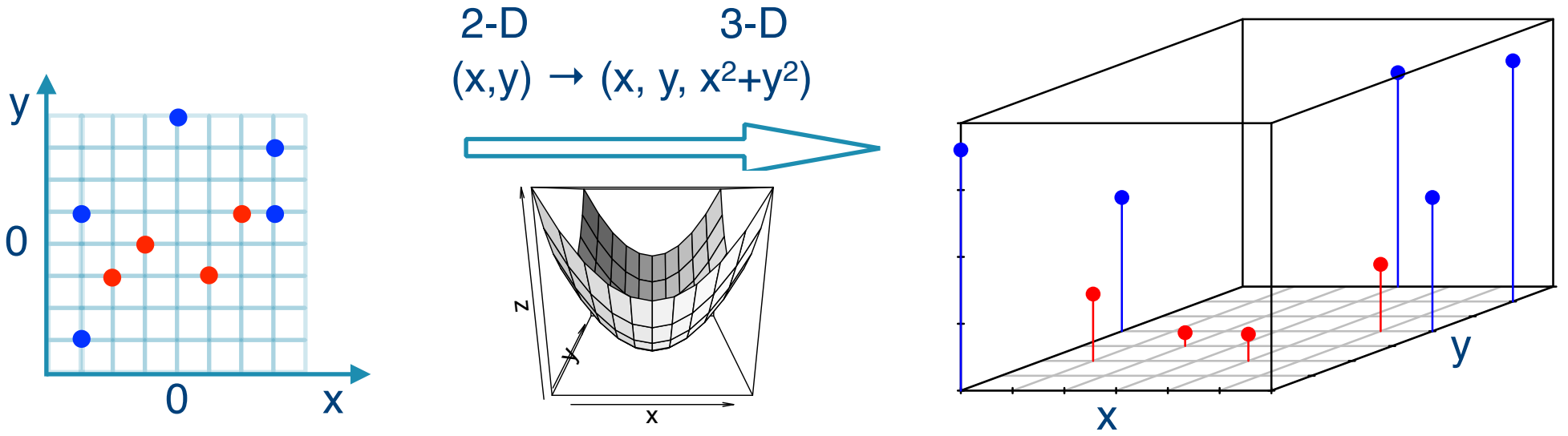


DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN

# Transformation to “feature space”

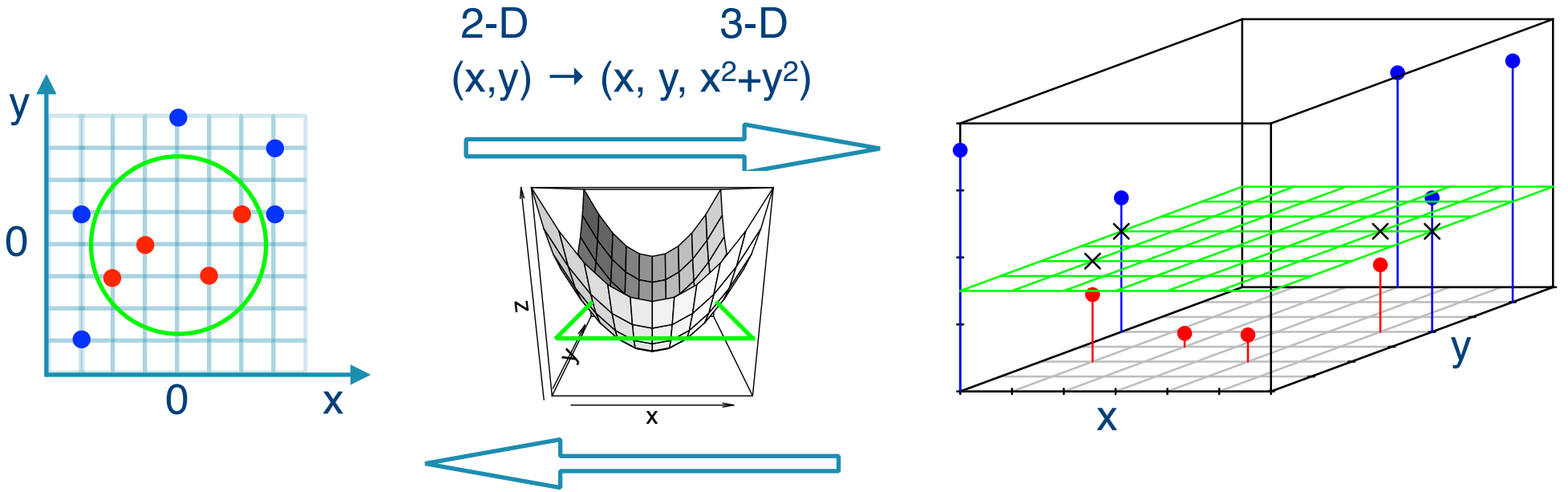
- Conceptually, SVMs learn non-linear boundaries by mapping the input space to a higher-dimensional “feature space”, constructing a max-margin linear boundary there, and mapping that back to the input space




(illustration - actual feature spaces have many more dimensions)

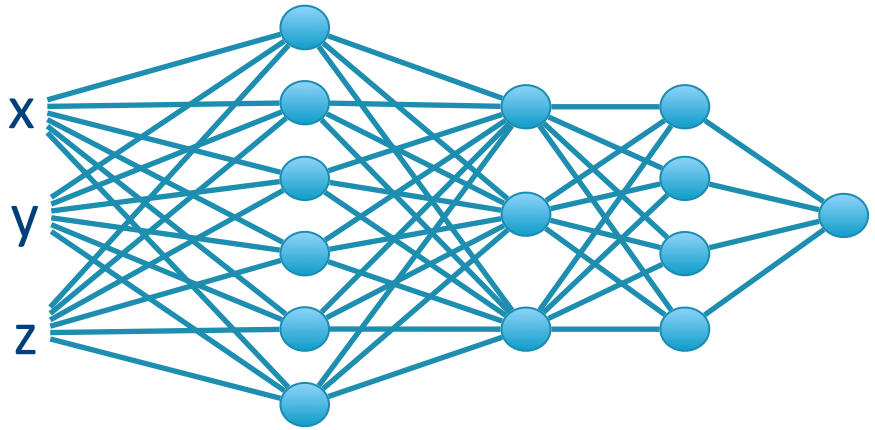
# Transformation to “feature space”

- Conceptually, SVMs learn non-linear boundaries by mapping the input space to a higher-dimensional “feature space”, constructing a max-margin linear boundary there, and mapping that back to the input space



# Method: Neural networks

- Very versatile format for numerical input-output functions
- ``Black-box" models: not easy to interpret
- Difficult to train, especially with many layers
  - Recent breakthrough: deep learning
  - But training requires much data and is very expensive

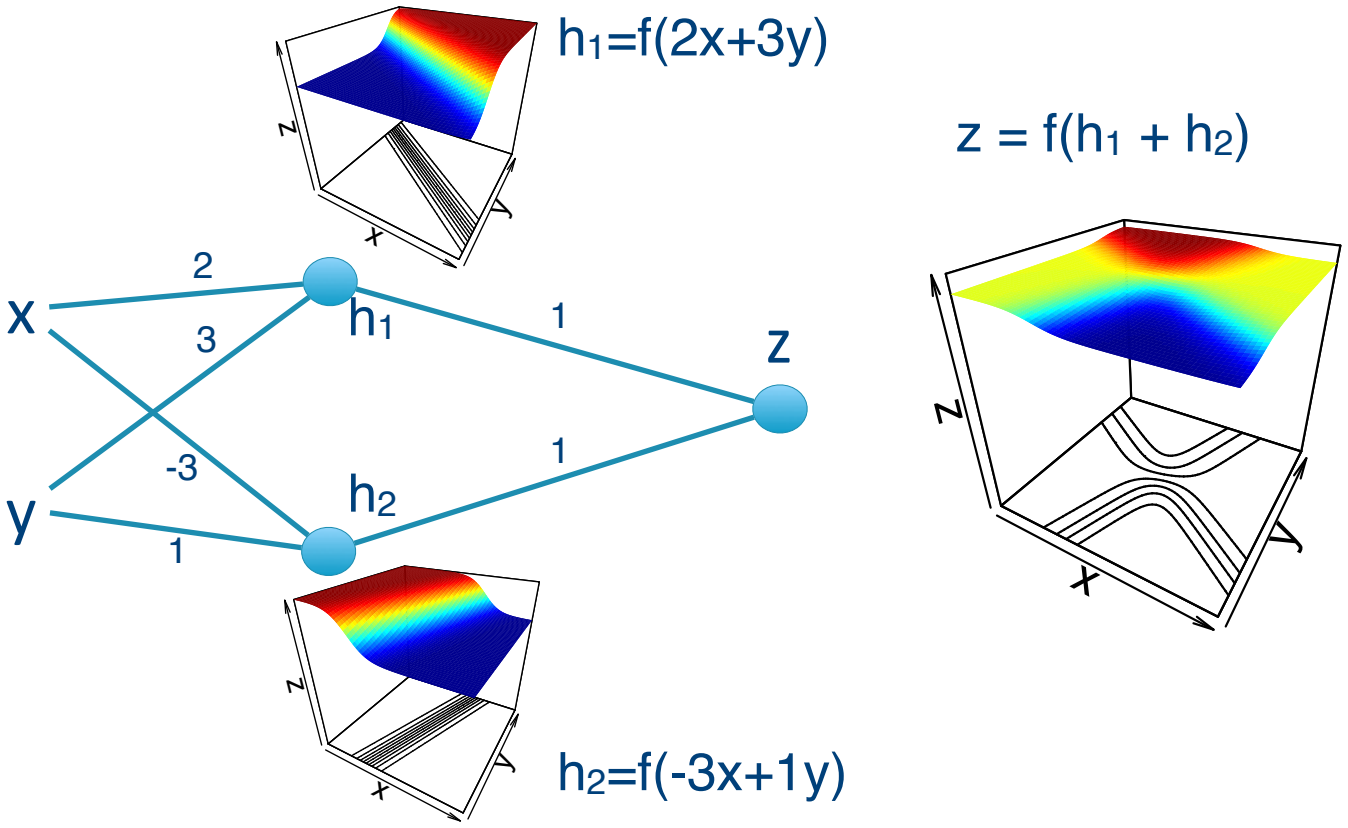



DTAI
Declaratieve Talen en Artificiële Intelligentie

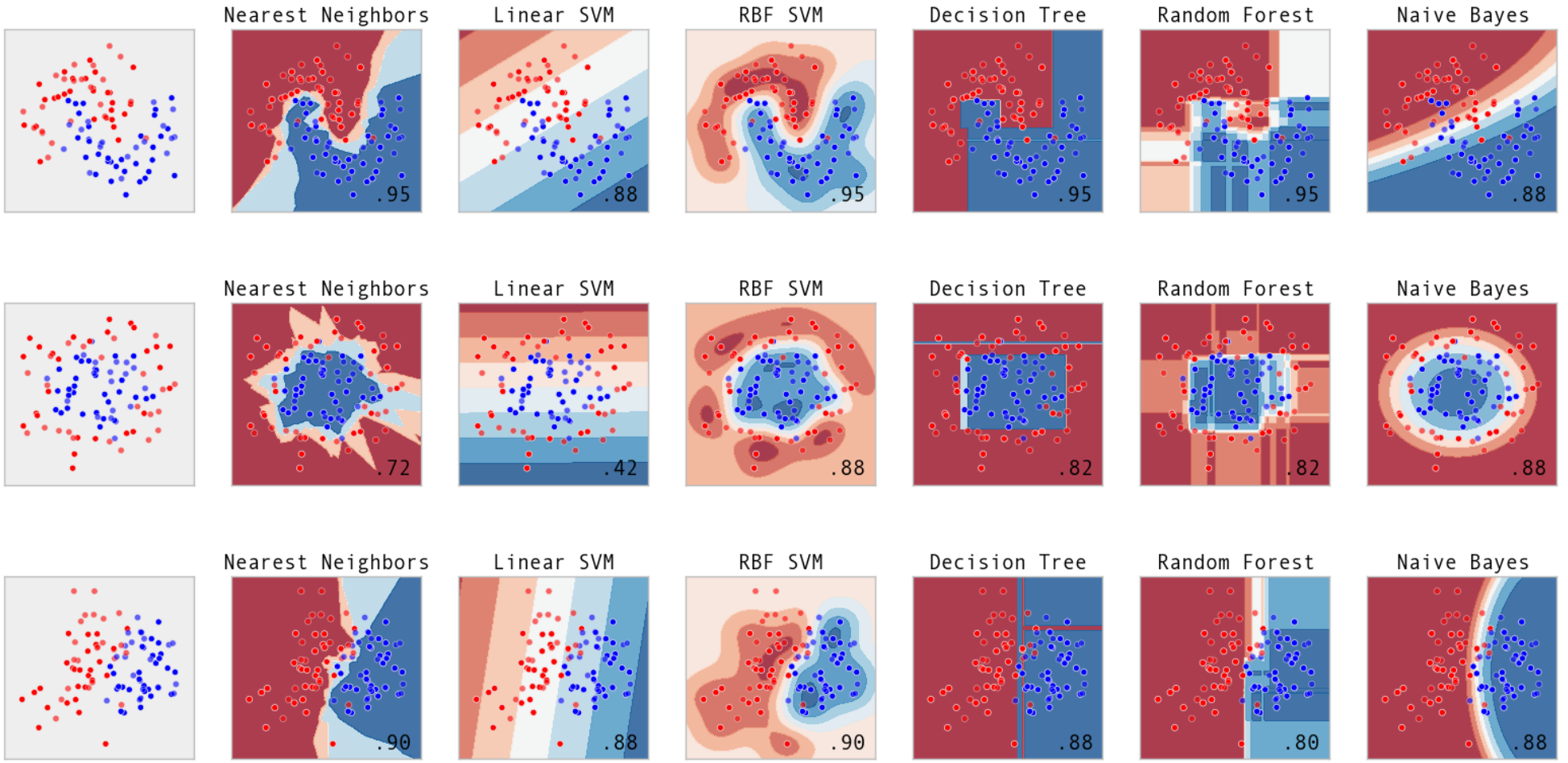
KU LEUVEN

# 2-layered feedforward network

- 2-layered perceptrons can construct non-linear separators
- The more “hidden nodes”, the more complex the separators can be



# Classifiers illustrated in 2-D



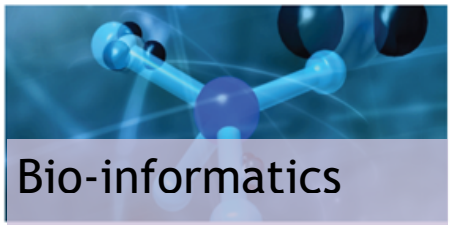

DTAI
Declaratieve Talen en Artificiële Intelligentie

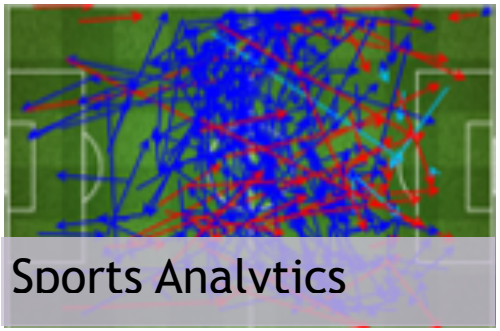
KU LEUVEN

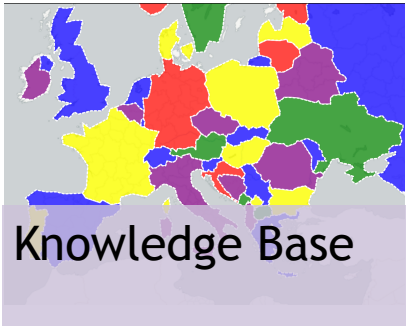
# To conclude…

- Data is now recognized as a major asset in many sectors of industry, science, society
- Large investments into collection, storage and analysis
- Data science technology is evolving fast: the pay-off of these investments will keep increasing

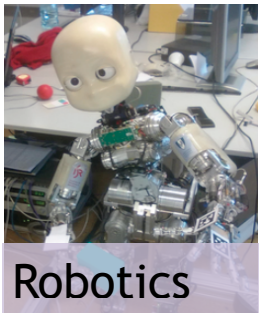
# Questions?



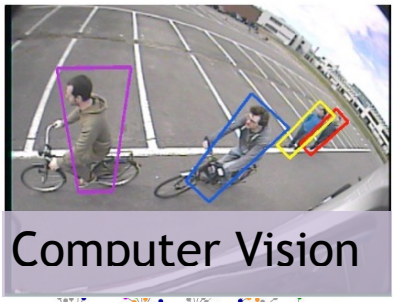



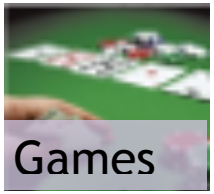


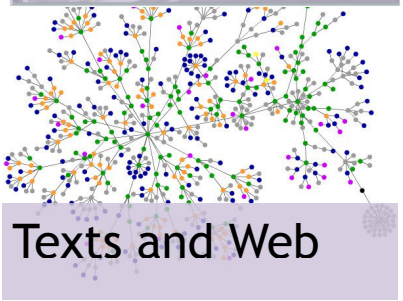












https://dtai.cs.kuleuven.be/topics

DTAI
Declaratieve Talen en Artificiële Intelligentie

KU LEUVEN